Topic 3: Sentiment Analysis

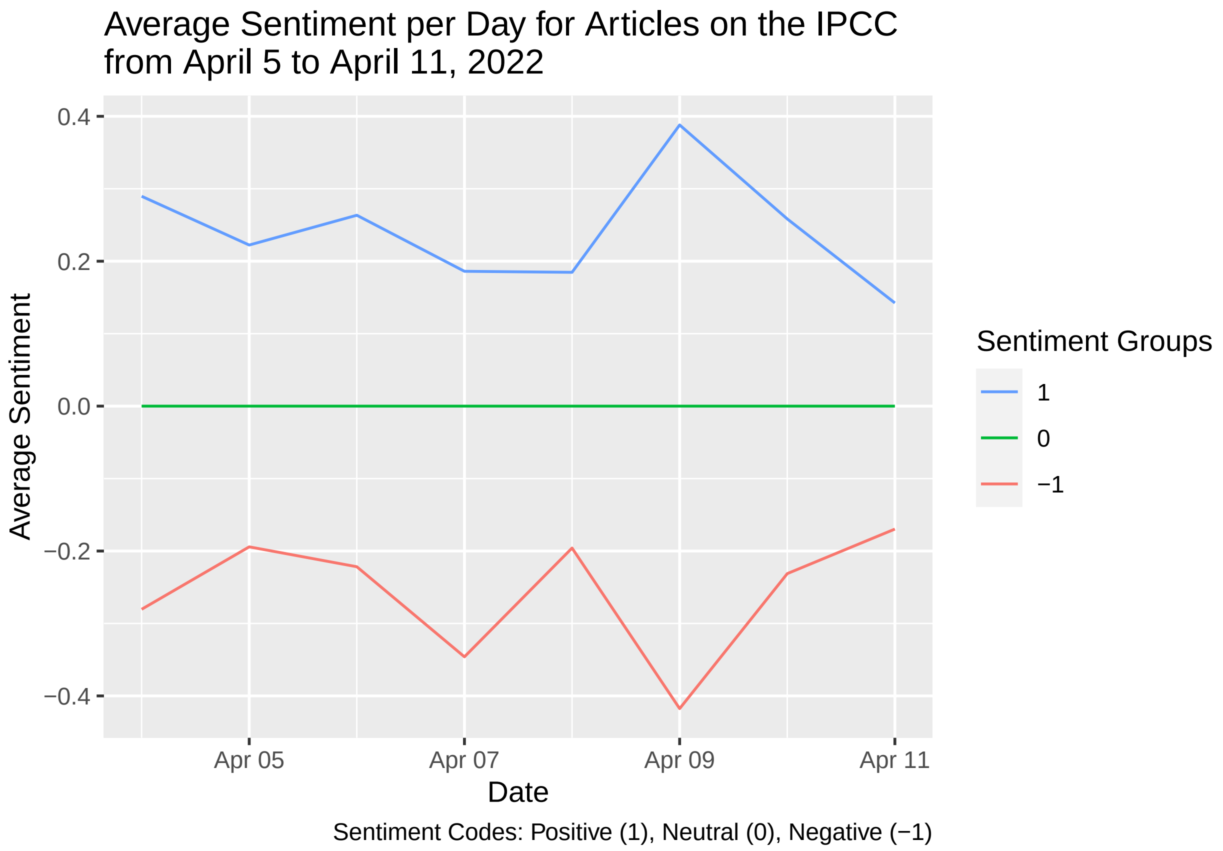
Clarissa Boyajian

2022-04-19

“IPCC” Nexis Uni data set

Use the “IPCC” Nexis Uni data set from class to create plot showing the average sentiment of headlines by day.

```
IPCC_files <- list.files(pattern = "Nexis_IPCC_Results.docx", path = here::here("data"),
                         full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

dat_IPCC <- lnt_read(IPCC_files) # Object of class 'LNT output'

# split LNT output class into three different dfs
meta_df_IPCC <- dat_IPCC@meta
articles_df_IPCC <- dat_IPCC@articles
paragraphs_df_IPCC <- dat_IPCC@paragraphs

dat2_IPCC <- data_frame(element_id = seq(1:length(meta_df_IPCC$Headline)),
                        Date = meta_df_IPCC$Date,
                        Headline = meta_df_IPCC$Headline)

# can we create a similar graph to Figure 3A from Froelich et al.?
mytext_IPCC <- get_sentences(dat2_IPCC$Headline)

# approximate the overall sentiment for a given text (scale -1 to 1)
# (attempts to correct for negation, context, etc.)
sent_IPCC <- sentiment(mytext_IPCC)

sent_df_IPCC <- inner_join(x = dat2_IPCC, y = sent_IPCC,
                           by = "element_id")

sentiment_IPCC <- sentiment_by(sent_df_IPCC$Headline)

# create plot
sent_df_IPCC %>%
  mutate(sentiment_groups = case_when(sentiment > 0 ~ "1",
                                      sentiment == 0 ~ "0",
                                      sentiment < 0 ~ "-1"),
         factor(sentiment_groups, levels = c(1, 0, -1))) %>%
  group_by(Date, sentiment_groups) %>%
  summarise(mean_sentiment = mean(sentiment)) %>%
  ggplot(aes(x = Date,
             y = mean_sentiment,
             color = sentiment_groups)) +
  geom_line(position = "dodge") +
  labs(col = "Sentiment Groups",
       y = "Average Sentiment",
       title = "Average Sentiment per Day for Articles on the IPCC \nfrom April 5 to April 11, 2022",
       caption = "Sentiment Codes: Positive (1), Neutral (0), Negative (-1)") +
  guides(color = guide_legend(reverse = TRUE))
```



"Heat Related Death" Nexis Uni data set

Access and wrangle data

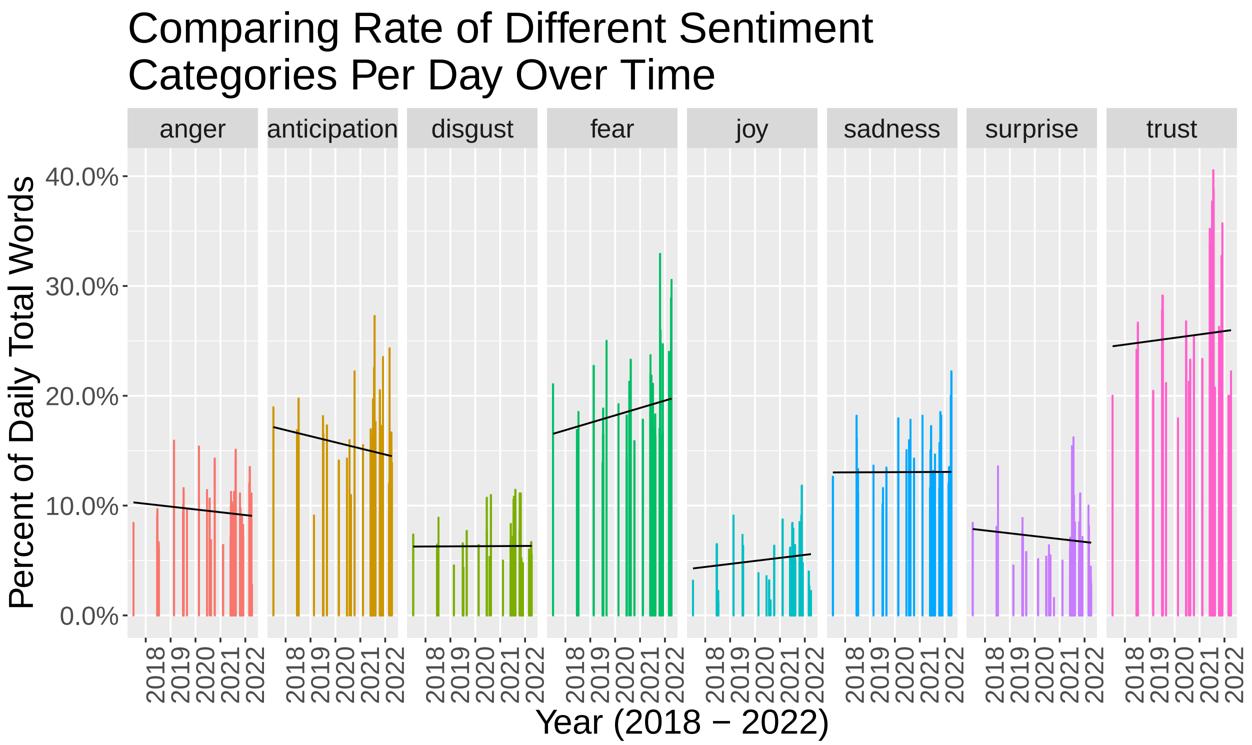
```
my_files_heat <- list.files(pattern = "Nexis_heat-related-death.docx",
                            path = here::here("data"),
                            full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

dat_heat <- lnt_read(my_files_heat) # Object of class 'LNT output'

# split LNT output class into three different dfs
meta_df_heat <- dat_heat@meta
articles_df_heat <- dat_heat@articles
paragraphs_df_heat <- dat_heat@paragraphs

dat2_heat <- data_frame(element_id = seq(1:length(meta_df_heat$Headline)),
                        Date = meta_df_heat$Date,
                        Headline = meta_df_heat$Headline)

paragraphs_dat_heat <- data_frame(element_id = paragraphs_df_heat$Art_ID,
                                  Text  = paragraphs_df_heat$Paragraph)

dat3_heat <- inner_join(dat2_heat, paragraphs_dat_heat, by = "element_id")
```

Clean data

```
# remove non-paragraphs from data
cleaned_data_heat <- dat3_heat %>%
  mutate(text_https = str_detect(string = dat3_heat$Text,
                                 pattern = "https",
                                 negate = TRUE),
         text_blank = str_detect(string = dat3_heat$Text,
                                  pattern = "^ $",
                                  negate = TRUE),
         text_length = str_length(string = dat3_heat$Text)) %>%
  filter(text_https == TRUE,
         text_blank == TRUE,
         text_length > 19)
# split paragraphs into individual words
cleaned_data_heat_words <- cleaned_data_heat  %>%
  select(!c(text_https, text_blank, text_length)) %>%
  unnest_tokens(output = word, input = Text, token = 'words')

# get 'nrc' sentiment lexicon from tidytext
nrc_sentiment <- get_sentiments('nrc')

# combine
cleaned_data_heat_sentiment_words <- cleaned_data_heat_words %>%
  # remove rows with stop words
  anti_join(stop_words, by = 'word') %>%
  # add emotion words
  inner_join(nrc_sentiment, by = 'word') %>%
  filter(!sentiment %in% c("negative", "positive"))
```

Create plot

```
# wrangle data for plot
data_heat_graph <- cleaned_data_heat_sentiment_words %>%
  group_by(Date, sentiment) %>%
  summarise(count = n()) %>%
  mutate(sum_count = sum(count))

# create plot of emotion words
ggplot(data = data_heat_graph,
       aes(x = Date,
           y = count / sum_count,
           color = sentiment)) +
  geom_col() +
  geom_smooth(method = lm, color = "black", size = 0.5, se = FALSE) +
  scale_y_continuous(labels = percent) +
  scale_x_date(date_labels = "%Y") +
  facet_grid(~sentiment) +
  theme(legend.position = "none",
        panel.grid.minor.x = element_blank(),
        axis.title = element_text(size = 20),
        axis.text.x = element_text(size = 15, angle = 90),
        axis.text.y = element_text(size = 15),
        plot.title = element_text(size = 25),
        strip.text = element_text(size = 15)) +
  labs(x = "Year (2018 - 2022)",
       y = "Percent of Daily Total Words",
       title = "Comparing Rate of Different Sentiment \nCategories Per Day Over Time")
```



Question

Over time, words that are associated with “trust”, “fear”, and “joy” have increased. Words associated with “anticipation”, “surprise”, and “anger” have decreased. And words associated with “disgust” and “sadness” have stayed roughly the same. Increase heat related deaths are an outcome of climate change so it makes sense to me that as climate change has become more accepted as fact over time, that the number of “trust” words would increase and the number of “surprise” and “anticipation” words would decrease. Additionally, heat related deaths and climate change are both becoming more prevalent as time goes on, so the increased use of “fear” words also makes sense.